Multi-View Registration for Feature-Poor Underwater Imagery

Nicholas Carlevaris-Bianco
Department of Electrical Engineering & Computer Science
University of Michigan
Ann Arbor, Michigan 48109
Email: carlevar@umich.edu

Ryan M. Eustice
Department of Naval Architecture & Marine Engineering
University of Michigan
Ann Arbor, Michigan 48109
Email: eustice@umich.edu

***Abstract*—This paper reports an algorithm for the registration of images with low overlap and low visual feature density—a typical characteristic of down-looking underwater imagery. Our algorithm exploits locally accurate temporal motion-priors and pairwise image correspondences to aggregate semi-rigid sets of sequential images. These sets are then used to search for visual correspondences across sets instead of between individual pairs of images. By simultaneously searching over multiple views, we increase the physical area seen by more than one image, effectively increasing the "field of view" of the image correspondence search. This increases the probability that the area viewed by both sets will contain enough visual features to register the sets. Our algorithm systematically reduces the uncertainty in the motion prior between the two sets resulting in a refined motion prior that is used to geometrically constrain the correspondence search between sets. This geometric constraint allows us to confidently identify local correspondences that would not be possible globally, further increasing our ability to register images in feature poor environments. We present results using a real-world ship hull inspection data set collected by an autonomous underwater vehicle.**

I. Introduction

Visually augmented navigation relies on the ability to find correspondences between images taken at different times in order to register the images and produce a constraint on vehicle motion. Unfortunately, underwater image capture is severely inhibited by the high attenuation of light in water and the need to conserve energy with strobed lighting. Because of these factors most underwater imagery is collected with a small field of view (FOV) (e.g., typically 45° FOV at $<$ 3–5 m altitude from subject) and low spatial overlap between images (e.g., $< 15\%$). In many unstructured underwater environments there may not be a sufficient number of visual features within the small region of overlap between individual pairs of images to reliably register them. This often occurs in areas of uniform texture such as patches of sand, mud or rock, and man made materials such as metal or concrete.

In this paper we present an algorithm that addresses the problems of low image overlap and low visual feature density by exploiting locally accurate temporal motion-priors and pairwise image correspondences to aggregate together semi-rigid sets of sequential images. We then search for visual correspondences between these sets instead of between individual pairs of images. By searching over multi-image sets we increase the physical size of the scene visible in more than one view, effectively increasing the "field of view" of the correspondence search. This increase in the area of image overlap increases the probability that the scene viewed by both sets will contain enough visual features to produce a constraint on the camera's motion. This provides a great advantage for feature matching in environments with a low density of interesting visual features. An example of a feature poor region is that of a ship hull as shown in Fig. 1. Between any pair of images there are at most four feature correspondences identified, too few for reliable pairwise registration. However, in aggregate across the two three-image sets there are a total of 10 correspondences, still few, but enough to attempt image registration. This illustrates why the proposed multi-view algorithm is successful—it is often able to aggregate together a sufficient number of correspondences even when the number of correspondences for any given pair of images is too few for pairwise registration.

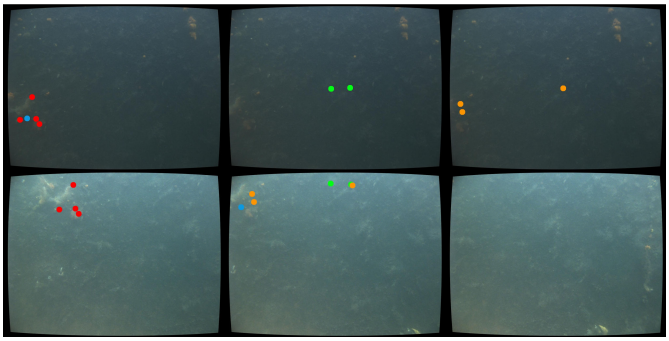


Fig. 1. Sample pair of image sets illustrating the utility of the multi-view correspondence search. The correspondences between images are marked with colored dots. A total of 10 correspondences are found between both image sets; however, between any *pair* of images there are at most 4 correspondences, too few to register. This illustrates why the proposed multi-view algorithm is successful: it is often able to aggregate together a sufficient number of correspondences even when the number of correspondences for any given pair of images is too few for pairwise registration.

It is important to emphasize that, though we present results for a mostly-planar ship hull data set, our algorithm makes no assumptions that the images in an aggregate set are related by a homography. Instead, our algorithm uses a multi-view geometry constraint and does not assume that the environment

(a) Reference image. (b) PCCS with a strong motion prior. (c) PCCS with a weak motion prior.

Fig. 2. An example of pose-constrained correspondence search (PCCS): (a) the first image with four sample feature points, (b) the 99.9% confidence ellipses where the corresponding features should lie based on a strong motion prior, (c) the 99.9% confidence ellipses from a weak motion prior. With a strong prior the reduction in the physical space where correspondences may be found greatly relaxes the requirements on the appearance-based feature matching as a feature only needs to be matched locally within the ellipse as opposed to globally over the whole image. However, as seen in (c), a weak motion prior will not adequately constrain the feature search.

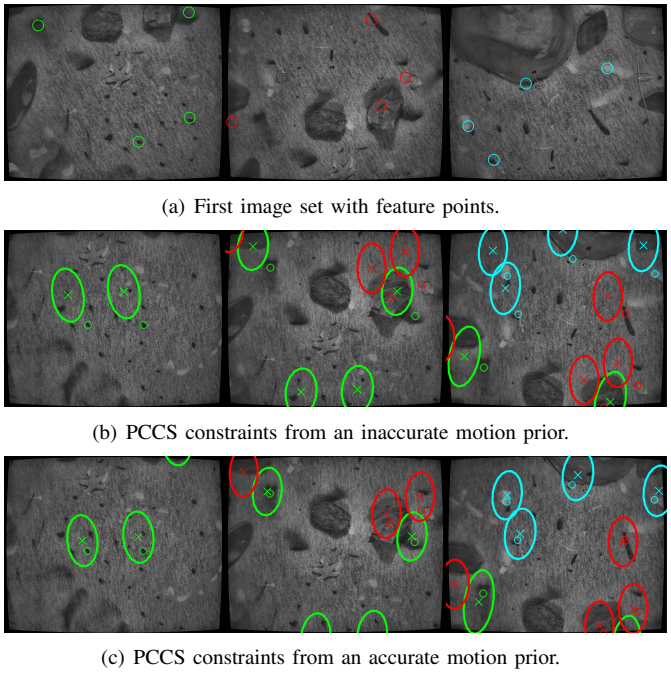
is planar, makes no homography-based approximations, and is equally valid for planar and fully 3D environments.

A. Background

It has been well established in the computer vision literature that epipolar geometry can be used to constrain the search space for feature correspondence [1]. Previous work by the authors [2] demonstrated that, on a robotic platform, the dead-reckoned (DR) pose prior between calibrated cameras and gross scene depth can be used to instantiate the epipolar geometry between frames and constrain correspondence establishment between a pair of images (Fig. 2).

This geometric constraint reduces the physical search space for a feature, which in turn reduces the required appearance-based uniqueness of the feature. This allows one to make confident matches locally that would not be possible globally, increasing the number of correspondences found. As an example, consider a small pebble on a sandy seafloor. If we were to attempt to locate that pebble in another image we most likely would not be able to, as there are many similar features on the seafloor. However, if we knew with a high degree of certainty that the pebble should exist only in a small region of the other image, then we would greatly increase the chances of identifying the same pebble in the other image.

In feature poor regions, this geometrically constrained search can provide a large advantage by allowing us to identify more subtle or repetitive visual features—for example, on an instrumented vehicle platform, this prior is typically available between sequential poses where the DR navigation is accurate. On the other hand, when the motion prior uncertainty is very large, this method no longer provides any advantage as the geometric constraints grow beyond the field of view (Fig. 2(c)). Hence, while the motion prior between temporally-sequential images *within* an image set will be strong, the motion prior *across* temporally-disjoint image sets may not be, greatly reducing (if not nullifying) the gains provided by the geometric constraint.



(a) First image set with feature points.

(b) PCCS constraints from an inaccurate motion prior.

(c) PCCS constraints from an accurate motion prior.

Fig. 3. Depiction of the discrete search used to refine the motion prior between sets. The first set is shown in (a) with four sample feature points for each image (circles). The geometric constraints (ellipses) shown in (b) are produced by an inaccurate estimate of the motion prior. We can see that none of the true correspondences, marked with small circles, lie within the correct geometric constraint ellipse. However, using a better estimate of the motion prior, (c), we find that all of the true correspondences now lie in their respective geometric constraint. A correspondence search using the motion prior in (b) would produce very few correspondences as its 99% confidence bound excludes all of the true matches. However, a search using the prior in (c) produce a very high number of matches. This response allows us to identify which discrete hypothesis, extracted from the original uncertainty in the motion prior, is closest to the true motion.

B. Methodology

In this paper, we propose a multi-hypothesis discrete search method to reduce the uncertainty in the motion prior between two sets. The discrete search divides the uncertainty in the initial motion prior into discrete hypotheses and then searches for the number of visual correspondences available for each hypothesis. For hypotheses that do not agree with the true motion between sets, there will be fewer putative visual correspondences. However, for hypotheses that agree closely with the true motion, we will see a very sharp increase in the number of correspondences found as the true correspondences will all fall within the geometric constraint. By repeating this process we can greatly reduce the uncertainty in the motion prior between sets, allowing us to geometrically constrain the visual correspondence search between sets. An example of this process is depicted in Fig. 3.

As advocated in [3] our proposed method for correspondence search first constrains the search region and then performs an appearance based search therein. This is in contrast with techniques, such as random sample consensus (RANSAC) [4], which first find appearance-based matches globally and then enforce geometric consistency. However,

unlike [3], where hypotheses are generated by sequentially searching feature by feature, we generate hypothesis by discretizing the highly uncertain dimensions of the pose prior. This allows us to start a search even when the pose prior uncertainty is initially very high.

Though the proposed algorithm is generically applicable in any unstructured or underwater environment, our particular motivating application for this work is autonomous ship hull inspection [5, 6]. During autonomous ship hull inspection an autonomous underwater vehicle (AUV) must accurately navigate along a ship's hull searching for damage or foreign objects. The collected images will almost invariably be of a flat, solid colored surface with very few remarkable features (Fig. 1). By increasing the effective field of view of the search, we increase the chances of finding a sufficient number of visual features to produce a constraint on the robot's pose.

It is important to note that the algorithm presented in this paper exists in the context of trajectory-oriented simultaneous localization and mapping (SLAM) [7] using vision as the primary sensing modality. In trajectory-oriented SLAM the robot maintains an estimate of its previous poses, referred to as the state estimate. Therefore, at a given time step, k, the estimated mean, μ, and covariance, Σ, for the current and all previous poses is available:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{p_k} \\ \boldsymbol{\mu}_{p_{k-1}} \\ \vdots \\ \boldsymbol{\mu}_{p_1} \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{p_k p_k} & \Sigma_{p_k p_{k-1}} & \cdots & \Sigma_{p_k p_1} \\ \Sigma_{p_{k-1} p_k} & \Sigma_{p_{k-1} p_{k-1}} & \cdots & \Sigma_{p_{k-1} p_1} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{p_1 p_k} & \Sigma_{p_1 p_{k-1}} & \cdots & \Sigma_{p_1 p_1} \end{bmatrix}$$

where each pose, $\mathbf{p}_i$, is a 6 degree of freedom (DOF) vector consisting of x, y, z translation and roll, r, pitch, p, and heading, h, attitude. Each pose is associated with the image collected at that location. Therefore, registering images provides a constraint on the motion of the AUV between the locations were the images were collected. In summary, the goal of trajectory-oriented visual SLAM is to produce the best estimate of the current and previous robot poses through the robot's DR navigation sensors and image registration. For a thorough overview of image registration and the related geometry we refer the reader to [8] and for an overview of SLAM we recommend [7, 9]. We also note that, although our algorithm does not depend on the particular selection of visual feature detector and descriptor, we have used the scale invariant feature transform (SIFT) [10] in our testing and to generate the figures in this paper.

The remainder of this paper is outlined as follows. In Section II we describe how prior information on the motion between two views can be used to geometrically constrain the visual correspondence search and in Section III we describe how these geometric constraints can be propagated through a set of images. In Section IV we describe how the initial uncertainty in the inter-set transform can be reduced using a discrete search and in Section V how the final pose constraint between sets can be estimated using bundle adjustment. In Section VI we present the results of the proposed algorithm as applied to a real-world ship hull inspection data set. Finally, in Section VII we offer some concluding remarks.

II. POSE-CONSTRAINED CORRESPONDENCE SEARCH

As previously mentioned in [2], it was shown that a-priori pose information can be used to provide a probabilistic geometric constraint on pixel locations where correspondences might be found between a pair of images. Pose-constrained correspondence search (PCCS) allows us to spatially restrict the search region in an image when establishing putative correspondences thereby reducing the required visual uniqueness of a feature. In other words, PCCS allows us to confidently identify correspondences that would not be possible using global appearance-based information only—since visual feature uniqueness no longer needs to be globally identifiable over the whole image, but rather it only needs to be locally identifiable within the geometrically constrained region.

In order to express the PCCS geometric constraint we start with two projective cameras with projection matrices defined as $\mathrm{P} = \mathrm{K}[\mathrm{I} \,|\, \mathbf{0}]$ and $\mathrm{P}' = \mathrm{K}[\mathrm{R} \,|\, \mathbf{t}]$, where R and $\mathbf{t}$ represent the rotation and translation between the two cameras and K is the camera calibration matrix [8]. If the distance from the camera to the scene point, Z, is known then the non-homogeneous point transfer mapping [2] can be used to project a point in the image coordinates of the first image into the image coordinates of the second:

$$\mathbf{u}' = \frac{\mathrm{H}_\infty \mathbf{u} + \mathrm{K}\mathbf{t}/Z}{\mathbf{H}_\infty^{3\top} \mathbf{u} + t_z/Z} \tag{1}$$

where $\mathrm{H}_\infty = \mathrm{KRK}^{-1}$ (often referred to as the infinite homography), $\mathbf{H}_\infty^{3\top}$ refers to the third row of H_∞, and t_z is the third element of $\mathbf{t}$.

With exact knowledge of all parameters the two-view point transfer mapping, (1), provides an exact mapping between image coordinates. In the robot's state estimate, however, R and $\mathbf{t}$ are only known up to a degree of uncertainty captured in the state covariance, Σ. Additionally, the scene depth (i.e., altitude) is instrumented with uncertainty. Therefore, instead of exactly projecting a point through (1) we instead find a first-order covariance ellipse in the second image's coordinate frame where we expect the point to lie based upon Σ.

To start we define our parameter vector, $\boldsymbol{\gamma}$, as

$$\boldsymbol{\gamma} = [\mathbf{t}, \boldsymbol{\Theta}, Z, u, v]^\top \tag{2}$$

where $\boldsymbol{\Theta}$ represents the roll, pitch, and yaw Euler angles comprising R. The parameter vector mean, $\boldsymbol{\mu}_{\boldsymbol{\gamma}}$, and covariance, $\Sigma_{\boldsymbol{\gamma}}$, are given by

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{t}} \\ \boldsymbol{\mu}_{\boldsymbol{\Theta}} \\ Z \\ u \\ v \end{bmatrix} \qquad \Sigma_{\boldsymbol{\gamma}} = \begin{bmatrix} \Sigma_{\mathbf{tt}} & \Sigma_{\mathbf{t}\boldsymbol{\Theta}} & 0 & 0 & 0 \\ \Sigma_{\boldsymbol{\Theta}\mathbf{t}} & \Sigma_{\boldsymbol{\Theta}\boldsymbol{\Theta}} & 0 & 0 & 0 \\ 0 & 0 & \sigma_Z^2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Here, the mean and covariance estimates of $\boldsymbol{\Theta}$ and $\mathbf{t}$ are extracted from the current state estimate, Z and σ_Z represent the scene depth parameters as measured in the first camera's frame, and (u, v) describe the feature location in pixels in the

first image. In defining $\Sigma_{\boldsymbol{\gamma}}$ we employ the standard assumption that features are extracted with isotropic, independent, unit variance noise [8] when defining the Σ_{uv} sub-block. To obtain a first-order estimate of the uncertainty in the point transfer mapping between the images we compute

$$\boldsymbol{\mu}_{\mathbf{u}'} \approx (1)\big|_{\boldsymbol{\mu}_{\boldsymbol{\gamma}}} \tag{3}$$

$$\Sigma_{\mathbf{u}'} \approx \mathrm{J}\Sigma_{\boldsymbol{\gamma}}\mathrm{J}^{\top} \tag{4}$$

where $\boldsymbol{\mu}_{\mathbf{u}'}$ is the predicted point location of $\mathbf{u}$ in I_j, $\Sigma_{\mathbf{u}'}$ its first-order covariance, and $\mathrm{J} = \frac{\partial \mathbf{u}'}{\partial \boldsymbol{\gamma}}$ is the point transfer Jacobian.

We use this knowledge to restrict our correspondence search using a Mahalanobis distance test:

$$\left(\mathbf{u}' - \boldsymbol{\mu}_{\mathbf{u}'}\right)^{\top} \Sigma_{\mathbf{u}'}^{-1} \left(\mathbf{u}' - \boldsymbol{\mu}_{\mathbf{u}'}\right) < k^2 \tag{5}$$

where the threshold k^2 follows a χ^2_2 distribution. Knowing the distribution of the threshold k^2 we can set it such that we obtain a set level of confidence in our estimate of the location of the point in the second image. In the remainder of this paper k^2 is set such that we are 99.9% confident the true correspondence lies within the projected ellipse.

An example of PCCS is shown in Fig. 2 where 2(a) shows the original points in the first image and 2(b) and 2(c) show the 99.9% confidence elliptical search regions based upon the uncertainty in the parameter vector $\boldsymbol{\gamma}$. In 2(b) a strong motion prior greatly reduces the search space, while in 2(c) a weak prior does little to reduce the search space.

III. Propagating Pose Constraints Between Sets

Section II described how prior information on the camera motion between two images can be used to geometrically constrain visual feature correspondences between two images. We will now describe how we can propagate the camera motion and its uncertainty through aggregate sets of images to constrain the correspondence search between image sets.

Fig. 4 illustrates the problem with two sample image sets, S_1 and S_2, each with three images $[A\ B\ C]$ and $[A'\ B'\ C']$, respectively. The intra-set transforms (between images within a set) are known with low uncertainty because the images are collected sequentially in time with little accumulated DR error (represented with thick black lines in Fig. 4). In order to constrain the robot's motion we want to find the inter-set transform (between set origins), $R_{S_1S_2}\ t_{S_1S_2}$, represented with a dashed red line. Because a large amount of time may have elapsed between the collection of S_1 and S_2, our knowledge of the inter-set transform may be very uncertain. Note that within the inter-set transform only the error in x and y grows unbounded with time. Depth z, roll r, pitch p and heading h are all easily instrumented with bounded error on most AUVs.

In order to produce a constraint between the two sets, we want to predict the locations of the feature points of S_1 in S_2 using the intra-set and inter-set transforms. We seek to to produce a constraint only on the single transform between set origins and not between the multiple transforms between the individual set images. Therefore, we propagate pairwise

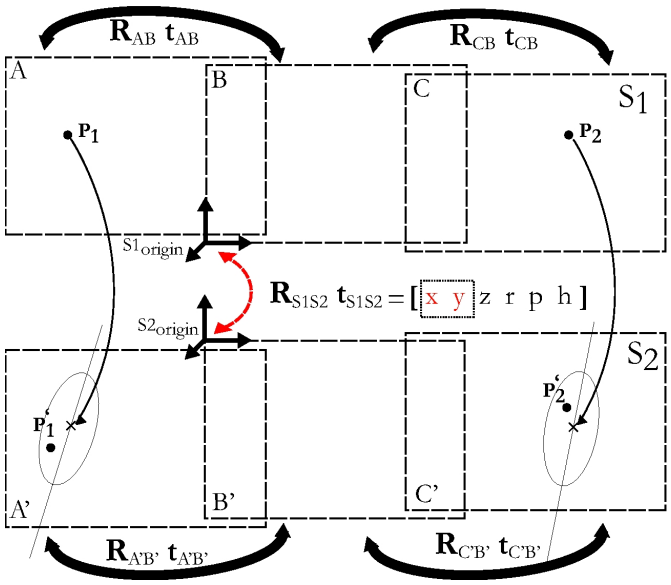


Fig. 4. Illustration of the inter-set matching problem with two sample image sets, S_1 and S_2, each with three images $[A\ B\ C]$ and $[A'\ B'\ C']$, respectively. The intra-set transforms (between images within a set) are known with low uncertainty and are represented with thick black lines. We want to find the inter-set transform, $R_{S_1S_2}\ t_{S_1S_2}$, represented with a dashed red line. The points P_1 and P_2 represent feature points in the first set while P'_1 and P'_2 represent their corresponding feature in the second set. The ellipses in the second set represent the 99.9% confidence region were we expect the points corresponding to P_1 and P_2 to exist given our estimate of the inter-set and intra-set transforms. Note that within the inter-set transform only the error in x and y grows unbounded with time. Depth z, roll r, pitch p and heading h are all easily instrumented with bounded error on most AUVs.

connections between S_1 and S_2 through the set origin images. Given n images per set we have n^2 pairwise connections between an image in S_1 and an image in S_2. Using the Smith, Self and Cheeseman notation [11] we propagate the transform between an image in S_1, I_i, and an image in S_2, I_j, as

$$X_{I_iI_j} = X_{I_iS_1} \oplus X_{S_1S_2} \oplus X_{S_2I_j} \tag{6}$$

where $\oplus$ is the head-to-tail operator. Note that for image pairs where I_i, I_j, or both are the origins of their respective sets the $X_{I_iS_1}$ and $X_{S_2I_j}$ transforms will be identity. Similarly, the covariance of the transformation is propagated as

$$\Sigma_{X_{I_iS_2}X_{I_iS_2}} = J^{\oplus}_{I_iS_2} \begin{bmatrix} \Sigma_{X_{I_iS_1}X_{I_iS_1}} & 0 \\ 0 & \Sigma_{X_{S_1S_2}X_{S_1S_2}} \end{bmatrix} J^{\oplus\top}_{I_iS_2} \tag{7}$$

$$\Sigma_{X_{I_i,I_j}X_{I_i,I_j}} = J^{\oplus}_{I_iI_j} \begin{bmatrix} \Sigma_{X_{I_iS_2}X_{I_iS_2}} & 0 \\ 0 & \Sigma_{X_{S_2I_j}X_{S_2I_j}} \end{bmatrix} J^{\oplus\top}_{I_iI_j} \tag{8}$$

where $J^{\oplus}_{I_iS_2}$ and $J^{\oplus}_{I_iI_j}$ are the Jacobians of the first and second head-to-tail operations, respectively.

With this method we can find $X_{I_iI_j}$ and $\Sigma_{X_{I_iI_j}X_{I_iI_j}}$ for any $i \in S_1$ and $j \in S_2$ allowing us to perform PCCS between any pair of images in the sets.

IV. Discrete Hypothesis Search to Reduce Inter-Set Uncertainty

If the initial uncertainty in the inter-set transform is too large, then we will not be able to adequately constrain the

location of possible features correspondences in the other set similar to that depicted in Fig. 2(c). Therefore, we wish to reduce the uncertainty in the motion prior between sets using a discrete hypothesis search (Fig. 3). We note that in most AUVs only the translation in x and y is instrumented with unbounded error. Depth, z, and attitude, r, p, and h, are all typically instrumented with bounded error. Therefore, when performing the discrete search we only need to search over x and y.

Given our current mean estimate of x and y, $\mu_{x,y}$, and their covariance, $\Sigma_{x,y}$, we seek to find a more accurate estimate, $\hat{\mu}_{x,y}$ and $\hat{\Sigma}_{x,y}$. This process is illustrated in Fig. 5. In 5(a) we start with the state estimate mean $\mu_{x,y}$ equal to the refined mean $\hat{\mu}_{x,y}$ and divide the translation uncertainty, $\Sigma_{x,y}$, into four hypotheses along its principle axes. Mathematically this division is performed as follows. First, we define the set pair state vector, $\boldsymbol{\gamma}$, as

$$\boldsymbol{\gamma} = [\mathbf{s}_1, \mathbf{s}_2, x, y, \boldsymbol{\zeta}]^\top \tag{9}$$

where the first and second intra-set transforms are represented by $\mathbf{s}_1$ and $\mathbf{s}_2$, respectively, and the inter-set transform is broken up into the unbounded components x and y and the bounded depth and attitude parameters, $\boldsymbol{\zeta} = [z, r, p, h]$. Next, we force the x and y components of the transform between S_i and S_j to be independent from the other pose parameters by forcing their correlations to zero:

$$\Sigma = \begin{bmatrix} \Sigma_{s_1,s_1} & \Sigma_{s_1,s_2} & 0 & 0 & \Sigma_{s_1,\zeta} \\ \Sigma_{s_2,s_1} & \Sigma_{s_2,s_2} & 0 & 0 & \Sigma_{s_2,\zeta} \\ 0 & 0 & \sigma_x^2 & \sigma_{x,y} & 0 \\ 0 & 0 & \sigma_{x,y} & \sigma_y^2 & 0 \\ \Sigma_{\zeta,s_1} & \Sigma_{\zeta,s_2} & 0 & 0 & \Sigma_{\zeta,\zeta} \end{bmatrix}. \tag{10}$$

We then perform a eigenvalue decomposition on the covariance of x and y to find the principal axes:

$$\Sigma_{x,y} = \begin{bmatrix} \sigma_x^2 & \sigma_{x,y} \\ \sigma_{x,y} & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^{-1}, \tag{11}$$

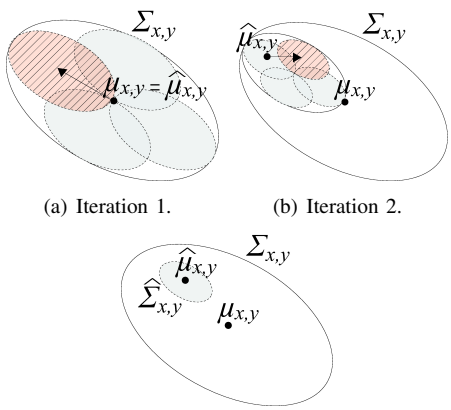
where V_1 and V_2 are the eigenvectors and σ_1^2 and σ_2^2 are the eigenvalues of $\Sigma_{x,y}$. The four divisions (the refined motion prior hypotheses) of the original covariance can then be found with

$$\hat{\mu}_{x,y} = \begin{cases} \mu_{x,y} \pm \frac{1}{2} V_1 \sqrt{\sigma_1^2 k^2} \\ \mu_{x,y} \pm \frac{1}{2} V_2 \sqrt{\sigma_2^2 k^2} \end{cases} \tag{12}$$

where k^2 is the desired χ_2^2 confidence level as discussed in Section II. The covariance for each of these four hypotheses is

$$\hat{\Sigma}_{x,y} = \Sigma_{x,y}/4. \tag{13}$$

Knowing the mean and covariance of the four hypotheses we then perform PCCS for each one, counting the number of putative visual feature correspondences found. In example Fig. 5(a), the upper left hypothesis (red with stripes) is found to produce the highest number of feature correspondences between the sets. We then repeat the process by again subdividing the hypothesis with the highest number of correspondences



(a) Iteration 1. (b) Iteration 2.

(c) Final refined estimate.

Fig. 5. Illustration of the discrete hypothesis search when x,y translational variance is high. In the first iteration, (a), the original 99.9% confidence ellipse is divided into hypotheses along its principle axes. The upper left hypothesis, red with stripes, is found to produce the highest number of feature correspondences and the estimate of x and y, $\hat{\mu}_{x,y}$ is moved to its center. The process is repeated in the second iteration, (b). Finally, after crossing a minimum covariance threshold or after seeing no further improvement in the number of correspondences found, a final refined estimate is produced, (c).

as shown in 5(b). The process repeats until the number of correspondences is no longer increasing, the magnitude of $\hat{\Sigma}_{x,y}$ is reduced below a threshold, or a maximum number of iterations is reached. In 5(c) the refined covariance, $\hat{\Sigma}_{x,y}$, is sufficiently small and $\hat{\mu}_{x,y}$ is accepted as the refined mean along with its maximal putative correspondence set of spatially constrained matches between the images in S_1 and S_2.

V. Estimating the Inter-Set Transform

In this section we describe how we refine our estimate of the inter-set transform, $R_{S_1,S_2}t_{S_1,S_2}$, where R_{S_1,S_2} is the rotation and t_{S_1,S_2} is the translation between the sets. Note that we parametrize $R_{S_1,S_2}t_{S_1,S_2}$ as a 6-vector, $X_{S1,S2} = [x, y, z, r, p, h]$, where x, y, z represent the translation and r, p, h, are the Euler angles of the rotation. Our proposed method is outlined in Algorithm 1.

Algorithm 1 Process Overview

```
Given S_i, S_j, μ, and Σ
if LargeXYUncertainty(Σ_{x,y}) then
    (μ̂_{x,y}, Σ̂_{x,y}) = DiscreteSearch(μ_{x,y}, Σ_{x,y})
else
    μ̂_{x,y} = μ_{x,y}
    Σ̂_{x,y} = Σ_{x,y}
end if
X = FindCorrespondences(μ̂_{x,y}, Σ̂_{x,y}, ...)
X = Triangulate3DPoints(μ̂_{x,y}, Σ̂_{x,y}, X, ...)
R̂t_{S_i,S_j} = BundleAdjustment(μ̂_{x,y}, Σ̂_{x,y}, X, X, ...)
```

Given two sets of sequential images that may overlap, S_i and S_j, the current state estimate, μ, and its covariance, Σ, we first check if the initial uncertainty in the inter-set transform exceeds a preset threshold. If it does we try to reduce the

uncertainty in the unbounded inter-set transform parameters as described in Section IV.

Using the refined inter-set transform, $\hat{\mu}_{x,y}$ and $\hat{\Sigma}_{x,y}$, and the intra-set transforms from the robot's current state estimate, we then find appearance-based feature correspondences, $\mathbf{x}$, between the images in S_1 and S_2 using the geometric constraints as described in Sections II and III. In the final step we calculate a refined translation between sets, $\hat{\mu}_{x,y}$ and $\hat{\Sigma}_{x,y}$, using a robust bundle adjustment [12] optimization. Bundle adjustment produces a maximum likelihood estimate (MLE) of the camera motion and the 3D scene structure. From this estimate of camera motion we can extract the MLE of the inter-set transform, $\hat{X}_{S_1 S_2}$.

Even though geometric constraints are enforced during the feature matching process, it is still possible that a small percentage of incorrect outlier correspondences will be present. These outliers, though few, can have a detrimental impact on the bundle adjustment if squared error is used as the criteria for optimization. Therefore, we instead use the Huber m-estimator [13] for the bundle adjustment optimization, which weighs small errors using a squared measure but then reduces the weight as the error increases, thereby reducing the effect of outliers. Another common approach would be to use RANSAC [4] *prior to optimization* in order to remove outliers. However, the number of pairwise correspondences between any two inter-set pairs may be *too few* to even attempt robustly fitting a pairwise motion model (Fig. 1). Hence, this is why we elect to use a robust m-estimator bundle adjustment framework.

It is also important to note that in addition to the feature correspondences found between sets we also include any features previously found to be inliers during intra-set pairwise image registration. This provides additional information for the bundle adjustment. Also, because the intra-set transformations are known with much less uncertainty than the inter-set transformation, we enforce a navigation prior during optimization that penalizes the modification of the intra-set transforms. This encourages the optimization to first modify the more uncertain transform between sets before adjusting the more precisely known transforms between images in a set.

We implement our navigation prior as described in [14] where the navigation prior is simply an additional measurement in the optimization parameter vector representing the Mahalanobis distance between the estimated intra-set transforms and their prior from the state estimate.

VI. Experimental Results

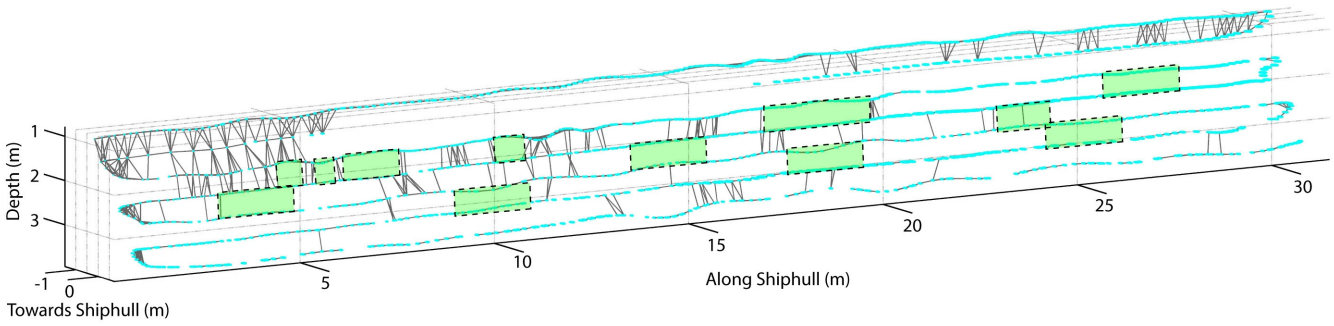
In order to verify the proposed algorithm we tested it over a real-world ship hull inspection dataset from the decommissioned aircraft carrier, USS Saratoga, collected at AUVFest2008 in Newport, Rhode Island using the Bluefin HAUV-2B [15] AUV. To do so, we used the output of the pairwise visually augmented navigation (VAN) algorithm [2] as a baseline for performance. We manually selected 324 image sets, varying in size from 3 to 5 images per set, and then grouped the sets into 162 inter-set pairs in regions where the pairwise algorithm was unable to find correspondence between images, but where we expected there to be image overlap based upon the trajectory of the robot. We then performed the proposed multi-view method over the manually selected sets to determine if it was possible to find links at these locations. Fig. 6 shows the results.

The original trajectory, after pairwise image registration is shown in 6(a). The cyan ellipses represent the trajectory estimate of the robot's motion and the gray links represent image correspondences found using the pairwise method. The 12 green boxes highlight regions where the pairwise method was unable to find any image correspondences. It was in these regions that the test sets for this paper were manually selected. The results of the set based registration are shown in 6(b). In this figure green lines represent successful new links, while red represents links that failed due to too few correspondences being found. In this experiment we used a minimum threshold of 10 inter-set correspondences. Finally, magenta lines represent links that failed because the bundle adjustment did not converge within 200 iterations of the Levenberg Marquardt (LM) algorithm. The results of the ship hull inspection dataset are summarized in Table I. Each successful pair represents an additional new pose constraint that the multi-view algorithm was able to provide. Overall the algorithm was able to identify a pose constraint in 79.63% of the proposed links. As previously mentioned, the image sets were selected with sizes varying from 3 to 5 images per set. The different sized image sets were distributed evenly throughout the robot's trajectory.

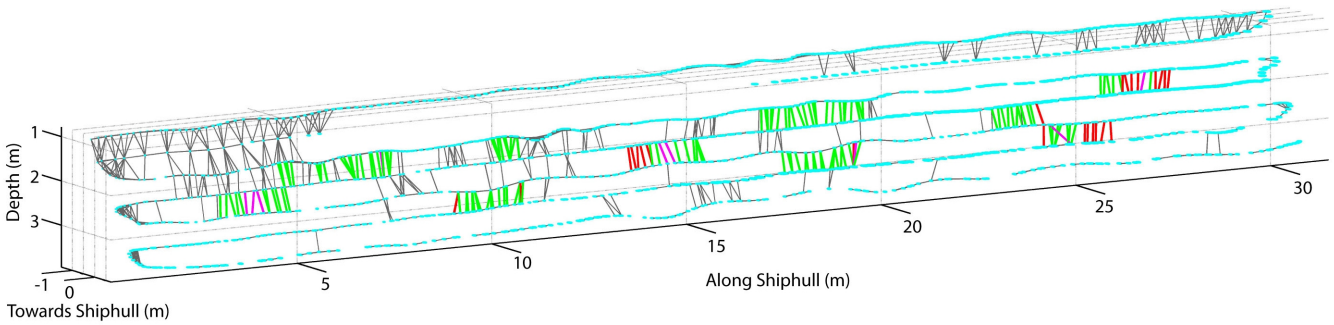
Comparing the links between image sets containing 3, 4, and 5 views we see that as we increase the number of images per set the number of successful links also increases. However, the increase in the number of views per set also results in an increase in the computational cost of inter-set matching. Therefore, the choice of image set size is a trade off between the improved probability of identifying a link, and the computational cost.

An additional consideration when selecting the image set size is the rate of accumulation of DR error. The ship hull inspection data set was collected using a highly accurate Doppler velocity log (DVL) velocity sensor that accumulates very little uncertainty over the set sizes used in our experiment. However, for less accurate DR sensors the position uncertainty within a set may grow to the point where adding additional images no long improves set registration as the intra-set uncertainty is too high to effectively constrain the visual correspondence search. Therefore one must also consider the rate of intra-set DR error when selecting the set size. In future work we plan to formulate a framework for automatically selecting the image set size based on the density of visual features in the scene, the computational cost of additional images, and the rate of accumulation of intra-set DR error.

In order to characterize the effects of the discrete hypothesis search described in Section IV, we performed the multi-view registration over the ship hull dataset with and without the discrete search enabled. When the discrete search was not enabled we simply used the current state estimate as the pose prior for



(a) Original VAN trajectory and camera-constraints.



(b) Additional links produced by our method.

Fig. 6. This figure depicts the results of our algorithm on a ship hull inspection dataset. In (a) the robot's trajectory is estimated using the pairwise VAN algorithm [2]. The cyan ellipses represent delayed-state robot poses and the gray lines represent pose constraints based on pairwise image registration. The green boxes highlight regions where the pairwise algorithm was unable to find sufficient correspondences for a constraint in the pose graph but where we expect there to be enough image overlap to find constraints. We manually selected 162 inter-set pairs from the green regions, varying in size from 3 to 5 images per set, and then used the proposed multi-view algorithm to attempt to find pose constraints between these sets. The results are shown in (b) where green lines represent successful pose constraints while red represents constraints that failed due to too few correspondences found and magenta lines represent constraints where the bundle adjustment failed to converge.

TABLE I
SUMMARY OF SHIP HULL DATASET DESCRIBING THE NUMBER OF LINKS ATTEMPTED AND FOUND USING THE PROPOSED MULTI-VIEW ALGORITHM

	Total	3 View	4 View	5 View
Attempted	162	54	54	54
Successful	129	41	43	45
Too Few Corrs.	24	11	7	6
No B.A. Convg.	9	2	4	3
% Successful	79.63%	75.93%	79.63%	83.33%

the PCCS. The number of additional correspondences found with the discrete search enabled is plotted for each pair of sets in Fig. 7. We can see from Fig. 7 that on average the discrete search finds an additional 6.66 correspondences per set. This is a significant number when our threshold for acceptance of a pair is on the order of 10 to 20 correspondences.

As described in Section IV the discrete search will only try to reduce the uncertainty in the initial inter-set transform if the state estimate uncertainty exceeds a threshold. Therefore, in Fig. 7, we have excluded the pairs where the algorithm found the initial pose uncertainty sufficiently small and therefore did not attempt to refine the pose prior.

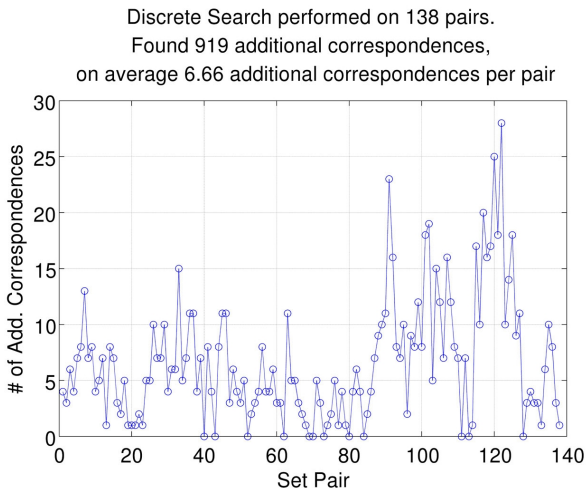


Fig. 7. This figure compares the number of correspondences found for each pair with and without the discrete search. For each pair of image sets the difference between the number of correspondences found with the discrete search and without is plotted. On average the discrete search finds 6.6 additional correspondences per set by reducing the uncertainty in the PCCS prior.

VII. Conclusion

This paper reported an algorithm to exploit locally accurate temporal motion priors to aggregate sets of sequential images for multi-view correspondence search. Simultaneously searching over multiple images increases the area of the visible scene, thereby increasing the probability of finding a sufficient number of visually interesting features in the overlapping image regions. This allows us to identify additional image-derived motion constraints that would otherwise be too weak to recover (due to a lack of feature correspondence density) in a pairwise registration framework. We demonstrated how our algorithm systematically reduces the uncertainty in non-informative intra-set motion priors via a greedy discrete search, and how the resulting refined motion prior can then be used to geometrically constrain the selection of intra-set putative correspondences. This geometric constraint allows us to confidently identify local correspondences that would otherwise not have been possible in a global matching sense—furthering our ability to register images in feature poor environments. Using a real-world ship hull inspection dataset, we showed how the multi-view algorithm discovered up to 83% more pose constraints than standard pairwise registration methods in feature-poor areas.

In future work we hope to explore how our method can be used for large loop closures in regions with a low density of visual features. We also hope to develop algorithms to automatically choose which sets of images to aggregate and which possible pairs of candidate sets to propose for link hypothesis. Finally, we also hope to explore the possibilities of applying our algorithm on terrestrial datasets where visual features are sparse or repetitive.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under award IIS-0746455 and by the Office of Naval Research (ONR) under award #N00014-07-1-0791.

References

[1] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," *Artif. Intell.*, vol. 78, pp. 87–119, October 1995.

[2] R. M. Eustice, O. Pizarro, and H. Singh, "Visually augmented navigation for autonomous underwater vehicles," *IEEE Journal of Oceanic Engineering*, vol. 33, no. 2, pp. 103–122, 2008.

[3] M. Chli and A. J. Davison, "Active Matching," in *Proceedings of the 10th European Conference on Computer Vision ECCV*, ser. Lecture Notes in Computer Science, D. Forsyth, P. Torr, and A. Zisserman, Eds., vol. 5302. Springer, 2008, pp. 72–85.

[4] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[5] A. Kim and R. M. Eustice, "Pose-graph visual SLAM with geometric model selection for autonomous underwater ship hull inspection," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, St. Louis, MO, 2009, pp. 1559–1565.

[6] R. M. Eustice, "Toward real-time visually augmented navigation for autonomous search and inspection of ship hulls and port facilities," in *Intl. Symposium on Technology and the Mine Problem*. Monterey, CA: Mine Warfare Association (MINWARA), 2008.

[7] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (SLAM): Part II." *IEEE Robotics and Automation Magazine*, vol. 13, no. 3, pp. 108–117, 2006.

[8] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2000.

[9] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robotics and Automation Magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[10] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," in *Autonomous Robot Vehicles*. Springer-Verlag, 1990, pp. 167–193.

[12] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment—a modern synthesis," in *Vision Algorithms: Theory and Practice*. Springer-Verlag, 2000, pp. 298–375.

[13] Z. Zhang, "Parameter estimation techniques: a tutorial with application to conic fitting," *Image and Vision Computing*, vol. 15, no. 1, pp. 59–76, 1997.

[14] O. Pizarro, "Large scale structure from motion for autonomous underwater vehicle surveys," Ph.D. dissertation, Massachusetts Institute of Technology / Woods Hole Oceanographic Institution Joint Program, 2004.

[15] J. Vaganay, M. Elkins, S. Willcox, F. Hover, R. Damus, S. Desset, J. Morash, and V. Polidoro, "Ship hull inspection by hull-relative navigation and control," in *Proceedings of the IEEE/MTS OCEANS Conference and Exhibition*, Washington, D.C., 2005, pp. 761–766.